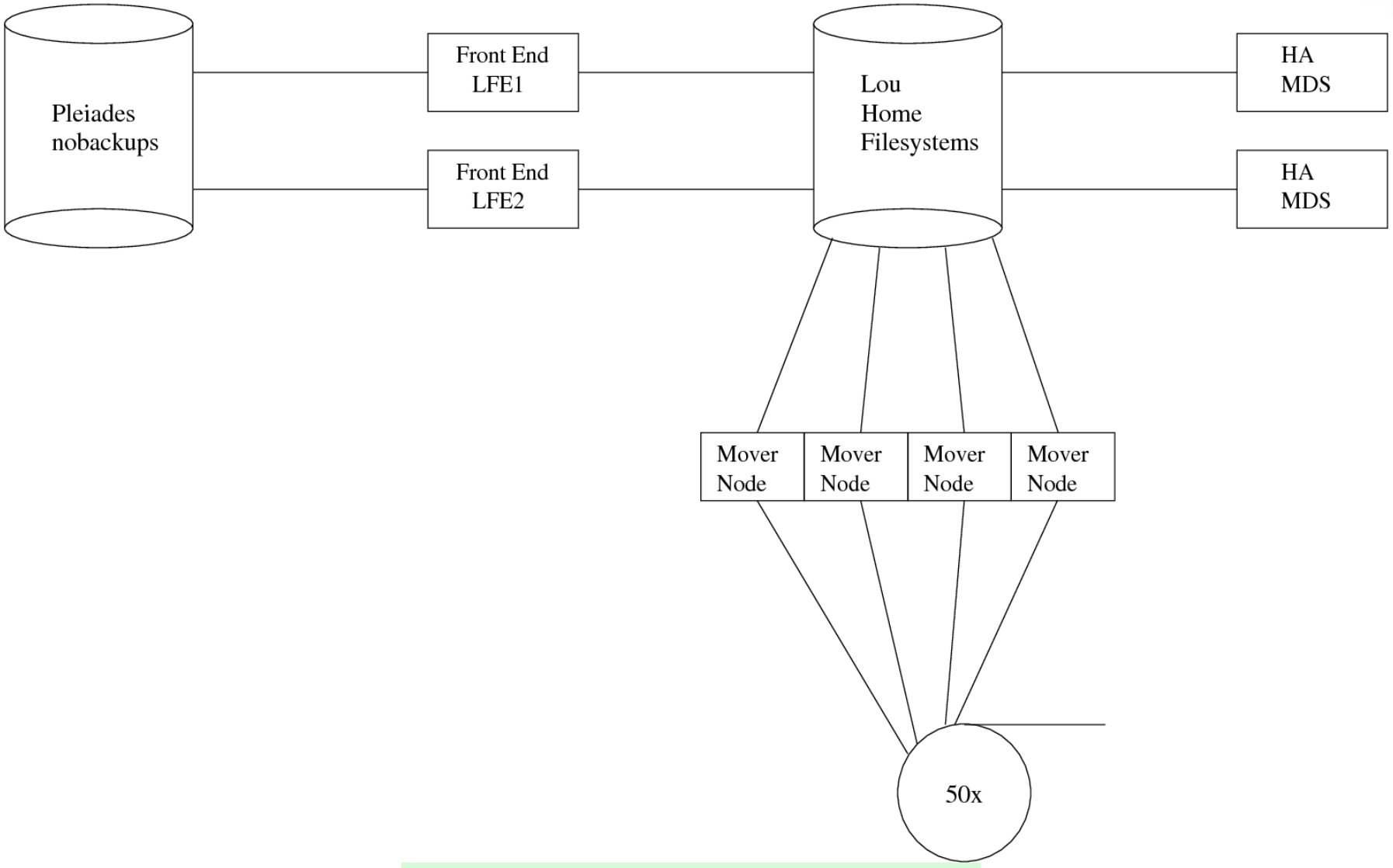
National Aeronautics and Space Administration

Effective Use of the New Lou2 Mass Storage Cluster

Jan 23, 2013

NASA Advanced Supercomputing Division

Diagram of Lou2 Cluster



Question? Use the Webex chat facility to ask the Host

Backing up your Data

- Don't think in terms of only backup/rm. They are independent functions
  - Back up the data when it has value and you want to keep it.
  - When you want to delete it from nobackup, verify it is saved to Lou and then delete it.
- We have lost entire nobackup filesystems
  - Each time multiple people say “I just lost six months work”.
  - Your workflow shouldn’t allow this.

A

Question? Use the Webex chat facility to ask the Host

Documentation

- The system is new. Best Practices are changing
  - URLs will change names over time
  - Search strings are included instead for the HECC
  - KnowledgeBase at http://www.nas.nasa.gov/hecc/
- More than one way to do things
  - Standard commands
  - Local, optimized commands

Question? Use the Webex chat facility to ask the Host

DMF Best Practices and Data Management

- Make sure to pre-fetch files with dmget
  - Run dmget on the same set of files you are going to work with and put it in the background (&)

    *dmget *.data &*

    *scp *.data matt@somwhere:*
  - KB search for “retrieval”

- Why do dmget. Worst case scenarios – observed
  - A tar done without pre-fetch was 1/3 done after a week; after a pre-fetch, it completed in two hours
  - 3 files were scp'ed. Two completed, but another user requested 100+ tapes and the third file took hours to return

DMF Best Practices and Data Management

- Use the --apparent-size of du to ensure you're not recalling too much data at once. What's excessive? >10TB? It depends.

  *du -sh --apparent-size dir_**

  dmfind dir_a dir_b -state OFL | dmget &

- Files written at the same time (within a few hours) tend to be on the same tapes.

- */usr/local/bin/dmfdu* will tell you the state of the recall, but it's slow for directories with many files.

  /usr/local/bin/dmfdu directory(s)

Tar Best Practices

- Many times tar files are too large – wastes resources
  - Worst observed case: 23TB tar file in a 30TB filesystem
  - Better to make more tar files in the data chunks you use
  - We try to limit files to 1TB
  - Use a shell loop to make multiple tar files
  - I can help!

- Make a table-of-contents file with *tar -tvf* or *mtar -tvf*
  - Use *mtar -tvf* or *mtar --print-hash -tvf*
  - Makes it easier to extract a subset of the tarfile
  - Gives "*ls -l*" style output; file dates are useful

Question? Use the Webex chat facility to ask the Host

Tar Best Practices

- Don't tar with gzip unless you have to for a slow WAN;
  It's 4x slower – our tape drives have built-in, line-rate HW compression so we compress the data on tape automatically

- Don't mv data from nobackup to Lou. Mv will be a copy/delete anyways but steals your chance to verify

Question? Use the Webex chat facility to ask the Host

Disk-to-Disk Copy (on Lou)

- Mtar or tar directly from nobackup to/from Lou home directory
  - This is the most preferred option from a systems POV
  - It's slower (for now), but it's a simple, one-step process

- Cd into nobackup to avoid extraneous paths in the tar

  cd /nobackupp1/mcary/datasets

  mtar -cf /u/mcary/datasets/set1a.tar set1a

- Mtar extract directly from Lou to nobackup

  mtar will lustre-stripe files written to nobackup

  cd /nobackupp1/mcary/datasets

  mtar -xf /u/mcary/datasets/set1a.tar

Question? Use the Webex chat facility to ask the Host

Disk-to-Disk Copy (cont)

- Shiftc will use whatever is most efficient – currently mcp
  - Same options as cp (mostly)

    cd /nobackupp1/mcary/datasets

    shiftc -rp set1a /u/mcary/datasets

- Shiftc will also Lustre stripe large files copied to nobackup
  - KB search “striping” and “shift”

- Cxfscp is an SGI-optimized cp command

  You have to stripe large files (>50GB) written to nobackup

  cxfscp -rplg (--bo/--bi)

  -rp Recursive, preserve timestamps, gid

  --bi, --bo buffer the nobackup I/O; --bi to Lou; --bo from Lou

  -l Treat links the same as cp

  -g Show transfer rate

Question? Use the Webex chat facility to ask the Host

Rsync (local or network)

- To Lou
  - Don't use --inplace or --checksum options; checksum will cause every file on Lou to be recalled from tape.
  - Do use -W option to work on whole files

- From Lou
  - Two rsyncs (or even three)

```
rsync --dry-run                 #Sanity check results
rsync --dry-run | dmget &       #Recall the needed files
rsync                           #Do the transfer
```


Question? Use the Webex chat facility to ask the Host

Local Network Copy

- "Lou" and "Lou2" are not hostnames
  - Scp/ssh can use these names
  - Bbftp/bbscp cannot by default

- Use the bridges or the pfes to transfer to Lou
  - The new pfes have 10GbE

Question? Use the Webex chat facility to ask the Host

Remote Network Copy

- To Lou
  - Use Secure Unattended Proxy (KB search “proxy” or “145”) for pre-authenticated, automated transfers or if both ends have two-factor. There was a Webinar last April if you want more information.

- From Lou
  - Use lfe2 if you have an existing hole in the remote firewall for Lou/Lou2.

Question? Use the Webex chat facility to ask the Host

Data Transfer from PBS job to Lou

- We don't allow transfers from compute nodes directly to Lou
  - Lou is not designed to handle 10K nodes
  - Jobs could stall, possibly for hours, waiting for files
  - Transfer rates would be poor

- Send a command to an intermediary (bridge/pfe)

  ssh -q bridge3 "shiftc -rp set1a lou2:/u/mcary/datasets"

Question? Use the Webex chat facility to ask the Host

Data Integrity

- Silent corruption does occur
  - *Shiftc --verify* does checksums at half the transfer rate

- Mtar has a feature to checksum an existing tar file

  cd /nobackupp1/mcary/datasets

  mtar -cf /u/mcary/datasets/set1a.tar set1a

  mtar -tf /u/mcary/datasets/set1a.tar --print-hash | md5sum -c

- If you just want to do a rough check that the tar file is correct

  du -sh --apparent-size /u/mcary/datasets/set1a.tar set1a

  find set1a | wc -l ; wc -l /u/mcary/datasets/set1a.tar.toc

Question? Use the Webex chat facility to ask the Host

Miscellany

- *cat /tmp/recallq* – a crude view of how busy tapes are for the last hour.

Timestamp	Files queued for recall	Tapes active or queued -Primary-	-All
15:52	0	6	22
15:54	1550	75	109

- If things are going slow, let us know

Question? Use the Webex chat facility to ask the Host

Futures

- The next version of shiftc (in a few months) will add
  - Faster tar – up to 15x faster than tar for disk-to-disk
  - Networked tar – copy a directory(s) to Lou as a tar file

- PBS-scheduled nodes to allow post-analysis of data on Lou.

- Sometime this year we hope to merge Lou1 and Lou2 and have only one mass storage system, Lou

Question? Use the Webex chat facility to ask the Host

Help

- We can help you with setting up or fine-tuning a workflow
  - support@nas.nasa.gov
  - 650-604-4444

- Who am I ?
  - Matt Cary
  - Mass Storage SysAdmin
  - matt.cary@nasa.gov
  - 650-604-4346

Question? Use the Webex chat facility to ask the Host